Immediate and Sustained Effects of Planning in a Problem-Solving Task*

By: Peter F. Delaney, K. Anders Ericsson, and Martin E. Knowles

Delaney, P. F., Ericsson, K. A., & Knowles, M. E. (2004). Immediate and sustained effects of planning in problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 1219-1234. Digital Object Identifier: 10.1037/0278-7393.30.6.1219

Abstract:
In 4 experiments, instructions to plan a task (water jugs) that normally produces little planning altered how participants solved the problems and resulted in enhanced learning and memory. Experiment 1 identified planning strategies that allowed participants to plan full solutions to water jugs problems. Experiment 2 showed that experience with planning led to better solutions even after planning was no longer required, whereas control participants showed little improvement. Experiments 3 and 4 showed that although the most recent planned solution could be recalled following a long filled retention interval, retroactive interference (RI) between successive problems resulted in much lower recall of earlier solutions. RI during plan generation could also explain participants' choice of depth-first planning strategies.

Article:
Planning ahead involves anticipating consequences of actions and results in memory representations that can range from complete step-by-step plans to strategic guides for future action. Although everyday plans are often supported by extensive world knowledge and integrated with goals and situational contexts (Friedman & Scholnick, 1997), unfamiliar problems may prove less conducive to planning. In fact, people usually do not plan ahead when solving unfamiliar problems (cf. Atwood, Masson, & Polson, 1980; Atwood & Polson, 1976; Best, 1980; Jeffries, Polson, Razran, & Atwood, 1977; Simon & Reed, 1976). One might therefore wonder how planning might be useful on unfamiliar problems and how encouraging planning could affect solutions.

One type of planning that has been observed on unfamiliar problems is subgoal-based planning (e.g., Altmann & Trafton, 2002; Anderson & Douglass, 2001; Klahr & Robinson, 1981; Simon, 1975; Ward & Allport, 1997). In some unfamiliar tasks (e.g., Tower of Hanoi), participants readily discover how to break the problem down into manageable subgoals during their initial analysis of the problem. In one recent study on subgoal-based planning, Ward and Allport (1997) instructed participants to solve a series of problems (using an easier variant of the Tower of Hanoi) in the fewest possible moves. Both a group that was specifically instructed to plan and a group that received no additional instructions spent time before making their first move to plan and subsequently solved the problems rather quickly. Planning time depended on the number of subgoals rather than

* Acknowledgement: This article is based in part on research reported in Peter F. Delaney's master's thesis and doctoral dissertation, both from Florida State University. We thank Joseph Anson, Alicia Bezmen, Sandra Nicole Bates, Chris Cardani, Tonya Carter, Taneikwe Grey, Kim Hrobek, Kimberly Jones, John Kim, Shannon Saunders, and David Southcott for their help in collecting and scoring portions of the data. We also thank Neil Charness, Jon Bailey, Fred Stephan, and John Shea, who served on Peter F. Delaney's thesis and dissertation committees and provided valuable suggestions. We gratefully acknowledge constructive comments on a draft of this article provided by Lili Sahakyan and Ira Fischler. This research was partially supported by Florida State College for Women/Conradi Endowment Fund of Florida State University Foundation and was prepared in part while K. Anders Ericsson was a Fellow at the Center for Advanced Study in the Behavioral Sciences. He would like to express his gratitude for the financial support provided by the John D. and Catherine T. MacArthur Foundation, Grant 32005-0, to the center.

the number of moves in the solutions, suggesting that participants used subgoal structures during planning. Subgoal-based planning may minimize memory demands, because the entire plan need not be represented in memory; the subgoals allow problem solvers to focus on a small part of the problem at any given time.

Our interest was in analyzing other types of planning behavior and identifying their potential benefits to problem solvers. Therefore, we sought a problem that lacks easily identifiable subgoals yet would still benefit from planning ahead. One problem that has these properties is a version of the water jugs problem (Atwood et al., 1980; Atwood & Polson, 1976). The water jugs problems are more challenging versions of the earlier problems used in mental set experiments (Luchins, 1942). The water jugs problems involve several jugs of different sizes. At the start of the problem, the first jug is usually full and the other jugs are empty. Throughout the problem, one cannot add or remove water from the system, so the total amount of water remains fixed. Moves consist of pouring water from one jug to another, stopping when the source jug is empty or the destination jug is full. The water levels that result from moves can be calculated. The goal is to match the water level of each jug to a predetermined goal.

The problem space for a typical problem is shown as Figure 1. The caption explains more about the problem-space diagram for readers unfamiliar with it. After each correct move toward the goal, there are usually four possible moves. Only one of those moves—and it is usually not obvious which one—moves closer to the goal. One of the incorrect moves is backward to a previous step along the goal path (and can be undone). Two incorrect moves (termed loopingmoves) irreversibly return to a state near the beginning of the problem and essentially force the problem solver to start over. For example, if the problem solver were at 2/3/3 (on the right side of the diagram), then pouring from C to B would move closer to the goal, resulting in 2/5/1. The move is reversible by then pouring from B to C. However, moving from B to A would result in State R, which is 5/0/3. This looping move could not be reversed by pouring from A to B again, as pouring from A to B at 5/0/3 would produce 0/5/3.

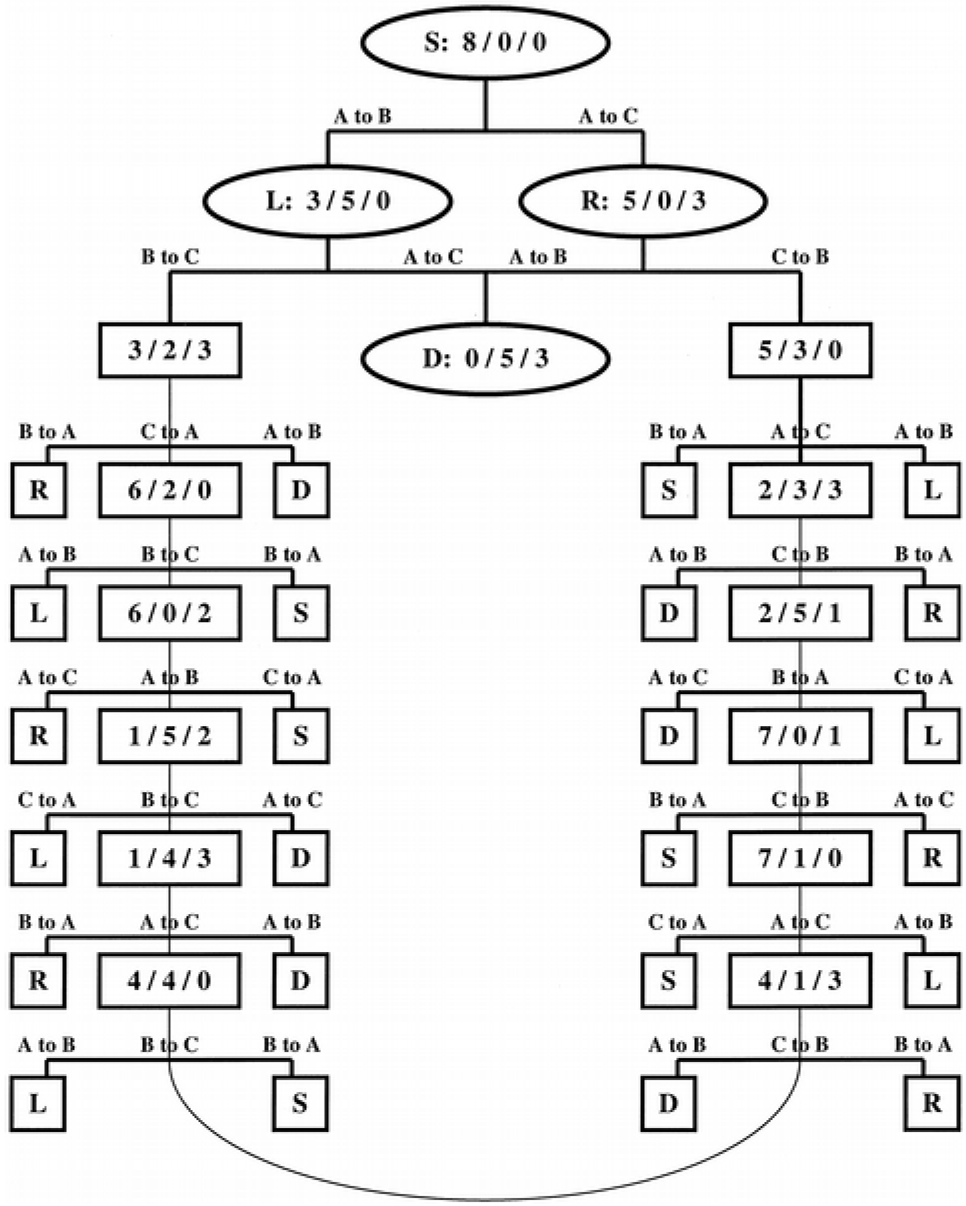

Figure 1. A problem-space diagram for a typical water jugs problem, showing the legal states and transitions between them. Legal states are represented by nodes, and moves are represented by paths between the nodes. The four oval nodes at the top of the diagram are given the special labels S (the starting state), L (left), R (right), and D (down) because irreversible looping moves always lead to one of these four states. The paths are labeled with the move that would be made to go from the higher state in the diagram to the lower state in the diagram (e.g., A to C indicates that one must pour from Jug A to Jug C). One can reverse moves that lead to a rectangle or oval node by making the opposite move (i.e., from A to B becomes from B to A). However, moves that lead to a square lead to the corresponding oval at the top of the diagram and cannot be reversed. Examples are given in the text

Atwood and Polson (1976) argued that mentally looking ahead one or two steps could often avoid the looping moves. Atwood and Polson's analysis of the problem structure also demonstrated that breaking the problem down into subgoals is nearly impossible at the start of the problem, because it is difficult to identify states along the solution path. Problem solvers might select an individual jug and try to match its contents to the goal before moving on to other jugs, but this strategy generally leads to the wrong move. To summarize, planning in the water jugs task could benefit from look-ahead planning, but it could not benefit from the structure provided by subgoals.

For many tasks, including water jugs and river-crossing problems like missionaries and cannibals, process models featuring even very simple look-ahead planning behavior outperform human problem solvers. Humans apparently consider only immediately available choices when selecting a move (Atwood et al., 1980; Atwood & Polson, 1976; Best, 1980; Jeffries et al., 1977; Simon & Reed, 1976). The process models explicitly proposed that limited short-term memory (STM) capacity prevented participants from using complex planning-ahead strategies. In Atwood and Polson's (1976) model, people first seek a move that has not been tried and that reduces the difference between the goal state and the current state. If none can be found, participants try to pick a move they have not tried before. Otherwise, they move randomly. More complex planning strategies would overload the model's severely limited STM, and long-term memory (LTM) serves only to avoid earlier visited states. Thus, planning-move sequences would prove impossible without extensive task experience, explaining why Atwood and Polson's participants did not plan. Consistent with their argument, Atwood et al. (1980) found that even presenting planning aids did not result in dramatic performance improvements.

Our first experiment was an exploratory test of several hypotheses concerning planning. First, we examined whether people would be constrained to planning very short sequences in the water jugs task as earlier process models suggest. We therefore instructed participants to mentally plan and then execute their solutions from memory without receiving any feedback about the outcomes of each move. Second, we used analyses of concurrent verbal reports to identify specific planning strategies adopted by our participants. We also used these reports to characterize how search strategies differed between participants instructed to plan and participants instructed only to solve the problems.

Our subsequent experiments investigated two different ways that people might learn from planning during problem solving. In Experiment 2 we explored whether requiring participants to solve some problems through planning would change how subsequent problems were solved after planning was no longer required. In Experiments 3 and 4 we investigated whether planning would enable participants to recall the solutions to multiple water jugs problems after a delay or after solving other problems. One reason to suspect that people could not recall multiple solutions is that the various moves and problem states in water jugs have been represented with a very similar encoding in memory. Considerable retroactive interference (RI) might be expected between alternative partial solutions or between successive problems and, hence, substantial forgetting of earlier solutions or attempts (Karat, 1982; Reed, Ernst, & Banerji, 1974, Experiment 2).

Experiment 1

In Experiment 1 we focused on whether instructing participants to plan would improve their problem-solving performance. To that end, participants solved water jugs problems under planning conditions and under control conditions. If the participants in our planning condition produced more efficient solutions (i.e., containing fewer moves) than the control condition, and those solutions were executed from memory (planning participants were not seeing the consequences of their moves to ensure that the problems were not solved again), then it would suggest that the solutions had been planned successfully. If Atwood and Polson's (1976) model were correct, STM should prevent effective planning and produce no benefit from the planning period.

In Experiment 1 we also explored the structure of planned solutions. Existing process models propose that people approach water jugs tasks by considering only the immediately available states (Atwood & Polson, 1976). The activities during planning might be no different than the activities during nonplanning, except insofar as they were done mentally and in advance. We expected few people to spontaneously plan in the water jugs task, in keeping with earlier research. However, we hoped to uncover evidence of complex planning strategies when planning was required.

Method

Participants

Participants were 57 Florida State University undergraduates who participated for course credit. They were screened for adequate arithmetic ability by a pretest consisting of four simple math problems. Seventeen participants were dismissed because they failed to solve at least three math problems after two attempts per problem. The remaining 40 participants completed both sessions.

Problems

During each of the two sessions, participants encountered one of the two problem sets (either Set A or Set B; see Table 1). Each problem set consisted of a practice problem and five water jugs problems, with an increasing minimum number of moves required to solve the problem throughout the series. Problems were shown so that participants solved easier problems before harder. Corresponding problems in Sets A and B were approximately equated for difficulty and solution length, and the set that was encountered first was counterbalanced.

Table 1
Problem Characteristics for Problems Used in All Four Experiments

Problem	Jug sizes	Initial state	Goal state	Solution length
		Experiment 1		
Practice A	5/4/3	5/0/0	2/3/0	2
Set A-1	12/5/3	12/0/0	6/3/3	3
Set A-2	8/5/3	8/0/0	6/0/2	4
Set A-3	10/5/4	10/0/0	7/0/3	5
Set A-4	10/7/3	10/0/0	9/0/1	6
Set A-5	11/7/3	11/0/0	3/5/3	8
Practice B	6/4/3	6/0/0	2/1/3	2
Set B-1	9/5/2	9/0/0	6/3/0	3
Set B-2	9/7/2	9/0/0	5/4/0	4
Set B-3	11/6/5	11/0/0	4/6/1	5
Set B-4	11/8/3	11/0/0	2/8/1	6
Set B-5	12/6/5	12/0/0	3/4/5	7
		Experiment 2		
Practice C	5/4/3	5/0/0	4/1/0	3
Set C-1	8/5/3	8/0/0	2/5/1	4
Set C-2	15/8/5	15/0/0	13/0/2	5
Set C-3	10/5/4	10/0/0	4/5/1	5
Set C-4	15/7/3	15/0/0	11/1/3	4
Practice D	6/4/3	6/0/0	4/0/2	3
Set D-1	9/4/3	9/0/0	3/4/2	4
Set D-2	9/5/4	9/0/0	3/5/1	5
Set D-3	13/7/5	13/0/0	10/0/3	5
Set D-4	14/9/4	14/0/0	9/1/4	4
		Experiments 3 and 4		
Practice-E1	6/4/3	6/0/0	4/0/2	3
Practice-E2	5/4/3	5/0/0	4/1/0	3
Set E-1	8/5/3	8/0/0	2/5/1	4
Set E-2	9/5/4	9/0/0	3/5/1	5
Set E-3	11/6/5	11/0/0	7/0/4	5

Problem Characteristics for Problems Used in All Four Experiments

Procedure

Testing occurred in two 1-hr sessions administered approximately a week apart. Participants were first trained to “think aloud” to ensure that they stayed on task and followed instructions (cf. Ericsson & Simon, 1993). Successful planning requires participants to perform the mental arithmetic when moving to calculate the contents of jugs, so participants also completed a prescreening test involving mental calculation. They then read the cover story (see the Appendix) and were shown the rules for water jugs and how to calculate the resulting water levels after moving.

During each session, participants attempted five water jugs problems in addition to an initial practice problem. A problem was considered correctly solved if the contents of the water jugs matched the goal state within the maximum time limit of 11 min per problem. The experimental interface was virtually identical to Atwood and

Polson's (1976) study.

Each participant completed one planning session and one control session (with order counterbalanced). During planning sessions, participants were instructed to make no moves until certain that they had generated the solution by planning (without writing anything down). When ready, they entered the moves into the computer. Throughout their solution, only the starting state of the problem was shown and no concurrent feedback was provided. If a wrong solution was entered, the participant was told to pour all the water back into the first jug and start over when they realized that their solution was incorrect (i.e., we gave no feedback unless the participant realized the solution was incorrect). During the control session, participants could begin moving at any time, and the computer updated the water levels after each move automatically. Thus, during the control session, participants could see the outcome of each move without having to calculate it.

Design

A 2 (experimental condition: planning instruction vs. control instruction) × 2 (test order: planning during first session and control in the second session vs. control during first session and planning in the second session) mixed factorial was used. Experimental condition was the within-subject factor.

Results and Discussion

Throughout Experiment 1, post hoc tests used all possible Bonferroni-corrected pairwise comparisons. Our primary hypotheses concerned differences in the number of moves participants made to solve the problems; therefore, Problem 1 was not analyzed further because virtually all participants solved it in the minimum number of moves in both sessions. The solution path was so short that many participants reported seeing the solution immediately, which defeated the purpose of planning. For the remaining four problems, we attempted to find a subset of problems that had similar solution rates in the planning and control conditions to avoid complicating selection-bias problems (e.g., because of speed-accuracy trade offs). To do so, a preliminary 2 (experimental condition) × 2 (test order) × 4 (problem type) analysis of variance (ANOVA) was conducted on solution accuracy. Table 2 reports descriptive statistics for proportion of correctly solved problems by problem type, 1 experimental condition, and test order. Because a significant Problem Type × Experimental Condition interaction emerged, $F(3, 114) = 3.65$, $MSE = 0.133$, $p < .05$, we conducted follow-up Experimental Condition × Test Order ANOVAs for each problem. The only significant effects were for Problem 5, where we observed significant effects of experimental condition, $F(1, 38) = 10.50$, $MSE = 0.144$, $p < .01$; test order, $F(1, 38) = 5.01$, $MSE = 0.202$, $p < .05$; and their interaction, $F(1, 38) = 10.50$, $MSE = 0.144$, $p < .01$. Post hoc tests revealed that accuracy was higher in the Session 2 control group than in all other groups. No other significant accuracy differences emerged. Therefore, we subsequently analyzed Problem Type 5 separately from the other problems.

Table 2
Mean Solution Rates for Experiment 1

	Problem type									
	1		2		3		4		5	
Session	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1										
Planning	.90	.31	.95	.22	.50	.51	.90	.31	.15	.37
Control	.90	.31	.90	.31	.85	.37	.85	.37	.20	.41
2										
Planning	.85	.37	.85	.37	.65	.49	.75	.44	.20	.41
Control	1.00	.00	1.00	.00	.70	.47	.85	.37	.70	.47

Mean Solution Rates for Experiment 1

Move Analyses

Atwood and Polson's (1976; Atwood et al., 1980) model would predict that our novice participants should have been unable to plan solutions in the water jugs task because of STM limitations. Observing substantial benefits of planning would therefore imply that our planning participants' behavior did not conform to their model's predictions. To test whether the planning period benefited the quality of solutions, we conducted a 2 (experimental condition) × 2 (test order) × 3 (problem type) mixed ANOVA on the number of extra moves (beyond the minimum) required to solve the problem, where experimental condition and problem type were within-subject factors. The number of moves, m, was log10(m + 1) transformed to reduce the violation of the normality assumption for ANOVA. Means are transformed back into normal units throughout this article to facilitate cross-experimental comparisons. The missing values that were due to unsolved problems were replaced with the grand mean of the transformed values.

Our critical prediction was that the control condition in either test order would require more moves to solve the problems than the planning condition in either test order (means relevant to this prediction are shown in Figure 2). The results confirmed the predicted main effect of experimental condition, $F(1, 38) = 42.03$, $MSE = 0.134$, $p < .001$, revealing that participants in the planning condition made fewer moves than participants in the control condition. There was also a main effect of test order, $F(1, 38) = 3.99$, $MSE = 0.143$, $p = .05$, indicating that participants on average made fewer moves when they received the control instructions after the planning instructions (3.57) than when they received the control instructions before the planning instruction (4.16).

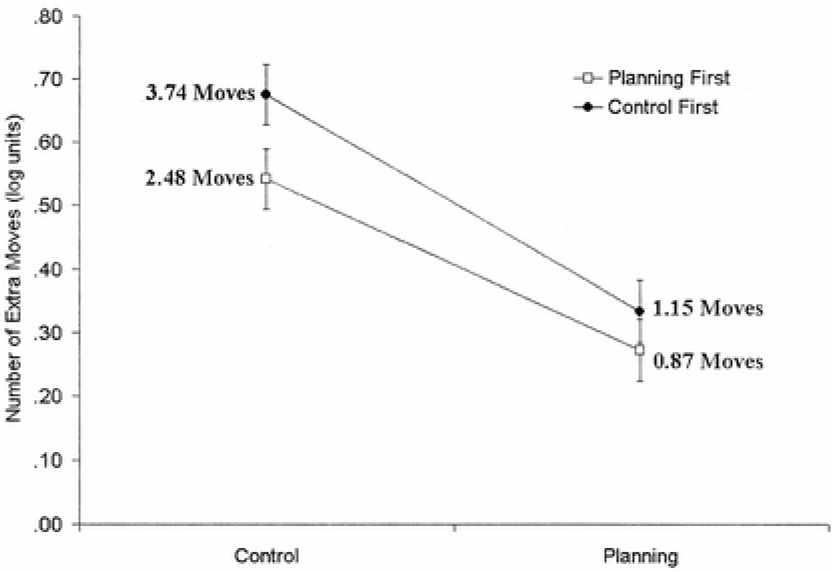

Figure 2. Number of moves committed beyond the minimum required to solve a problem as a function of experimental condition and order of conditions in Experiment 1. Axis values were log(x + 1) transformed. Error bars represent plus or minus standard errors

Last, a main effect of problem type emerged, reflecting variations in difficulty across problems, $F(2, 76) = 4.86$, $MSE = 0.157$, $p < .01$. Post hoc comparisons revealed that fewer extra moves were required for Problem Type 2 (2.3 moves) than Problem Types 3 or 4 (3.5 and 2.9 moves, respectively), which did not differ. No significant interactions emerged in this analysis.

Our most important result was that participants' solutions benefited from planning instructions. The executed solutions in the planning condition were virtually error free and were reliably shorter than the solutions in the control condition. The modal number of extra moves beyond the minimum in the planning group was 0 extra moves, even though participants received no feedback about the consequences of their moves. In comparison, the control condition produced less efficient solutions, despite receiving immediate feedback about the results of making moves and not having to keep track of the solution path they had taken.

Concurrent Reports: Structure of Planned and Unplanned Solutions

Atwood and Polson (1976) showed that most people look only at the current state and use simple heuristics to choose their moves. Planners might mentally use quite similar procedures. Alternatively, planning could produce qualitatively different solution processes. We examined concurrent verbal reports to address this issue.

Audiotapes from 20 participants' first sessions (10 in the planning group and 10 in the control group) were randomly selected for transcription, with the constraint that at least three or more problems were correctly solved. From the planning condition protocols, we extracted participants' verbalized moves. Fewer than 2% of verbalized moves could not be uniquely encoded and thus were not analyzed.

To address whether planners and control participants considered the same number of moves while searching for the solution, a 2 (experimental condition: planning vs. control) × 3 (problem type: 2, 3, and 4) ANOVA was conducted on the number of moves beyond the minimum solution. For control participants, the dependent measure was the actual number of moves made (because their solution processes are better reflected in their moves rather than verbalizations), whereas in the planning group it was the number of verbalized moves (including all missteps). Missing values were replaced with the grand mean. No differences emerged between the number of moves per problem considered by the planning group (4.23) and the number of moves made by the control group (4.57), resulting in no main effect and no interaction (both $Fs < 1$). There was only the expected main effect of problem type, reflecting difficulty differences across problems. Thus, planners seemingly did not produce more efficient search paths than control participants in terms of the number of considered moves.

That both types of solutions required similar numbers of moves need not imply that identical processes produced the moves. Atwood and Polson (1976) showed that participants usually generated many looping moves during their solutions. That is, the participants frequently made irreversible wrong moves that returned them to a state near the start of the problem (see Figure 1; an example is moving from A to B from 2/3/3, taking the solver to 0/5/3. This move cannot be reversed by moving from B to A, which would result in 5/0/3). Our next analysis examined whether planners and control participants made similar numbers of looping moves. Differences would likely indicate that different processes were used to generate the solutions in the planning and control groups. A 2 (experimental condition: planning vs. control) × 3 (problem type) ANOVA was computed on the number of looping moves. Planning instructions produced fewer looping moves per problem (.06) than control instructions (.53), $F(1, 18) = 9.80$, $MSE = 0.333$, $p <.01$. The three problems also differed, $F(2, 36) = 3.50$, $MSE = 0.400$, $p <.05$, with.10,.20, and.60 loops, respectively, for the three problems. The interaction approached but did not reach significance, $F(2, 36) = 2.67$, $p =.08$. Thus, although participants in the planning condition considered about the same number of moves, they almost never made looping moves, suggesting that qualitatively different processes produced the solutions in different conditions.

Why and how did planners make about the same number of total moves if they made fewer looping moves? We suspected that they used depth-first search, which involves following a path until no more acceptable moves can be found and then restarting from the starting state (by putting all the water back into the biggest jug; this could be done by pouring from B to A and C to A in the simulation). We explored differences in restarting by summing visits to the start state for each problem. A 2 (experimental condition) × 3 (problem type) ANOVA on the number of restarts showed that planning participants restarted the problem more often (1.50 restarts/problem) than the control group (0.33 restarts/problem), $F(1, 18) = 31.96$, $MSE = 0.639$, $p <.001$. Neither the main effect of problem type, $F(2, 36) = 1.27$, $MSE = 1.644$, nor the interaction was significant ($F < 1$). Thus, planners frequently restarted the problems, suggesting that they used depth-first search, whereas control participants did not. The depth-first search pattern is typical of planning in other domains such as chess (Charness, 1981) and may reflect the way that participants cope with working-memory limits. By constraining the number of moves considered by restarting often and using depth-first search, participants in the planning condition minimized the working-memory demands of planning. This perhaps explains how they were able to cope with the extreme working-memory demands identified by earlier researchers on the water jugs problem (Atwood et al., 1980; Atwood & Polson, 1976).

Concurrent Reports: Strategies on Planned and Unplanned Solutions

Most models of novice problem solving mainly rely on local information and heuristics to make decisions, not complex planning. Thus, our next set of analyses was concerned with the kinds of strategies that emerge under control instructions and planning instructions. We examined Session 1 concurrent protocols from the 3 problems from the same 10 participants in each condition in search of evidence of complex planning (for a total of 30 problems). Broadly speaking, by complex planning we refer to planning strategies that look beyond the current state to identify the best move.

For the planning group, several specific strategies emerged that could be classified as complex planning. The two most common strategies were as follows: (a) stating an arithmetic rule and then using those calculations to guide move selection (8/30 problems)—for example, 1 participant realized that Jug C held 2 liters of water and her goal in Jug B was 4 liters; therefore, she filled Jug C twice, pouring it each time into Jug B to get 4; and (b) verbalizing a plan to make a sequence of two or more moves to achieve a goal (5/30 problems). In each case this occurred when those two steps would reach the goal state. In addition, there were two relatively rare strategies, which were as follows: (a) setting higher order goals that were based on a sophisticated analysis of the problem (2/30 problems) and (b) using a strategy of filling the smallest jug repeatedly and pouring its contents into the middle jug (1/30 problems).

For the control group, complex planning was less common, and only the two most common planning strategies from the planning group were observed: (a) verbalizing a multistep plan involving the rules of arithmetic guiding move selection (1/30 problems) and (b) verbalizing two or more moves as a sequence (2/30 problems). In addition, there were 2 problems where the rules of arithmetic were used to guide move selection but did not result in progress toward the solution. However, 2 participants planned entire solutions before making any moves for several problems (like planning participants were instructed to do) or made substantial progress before beginning to take any moves (on a total of 4 problems). All of the cases of using the rules of arithmetic to guide solution (whether successful or not) involved participants who planned complete solutions.

Overall, of the 10 planning participants, 80% showed evidence in their verbal reports of some complex planning strategies. Using the same criteria, only 40% of the participants in the control group showed evidence of multistep planning (and some of these participants were actually solving the problems in a fashion similar to the planning group). To determine whether planners used complex planning strategies more often than control participants, the number of problems (0 to 3) where complex planning strategies were used was calculated for each participant and subjected to a t test. Planners were more likely to engage in complex planning strategies (1.30 problems/participant) than the control group (0.50 problems/participant), $t(18) = 1.99$, $p < .05$. Thus, planners discovered complex planning strategies more frequently than control participants did.

The Hardest Type of Problems

Because of the interaction of accuracy across sessions and experimental conditions for Problem Type 5, we analyzed it separately. Only 4 participants solved this type of problem in both sessions, so we could not capitalize on the within-subject design. The number of moves beyond the minimum was therefore analyzed for the 25 solutions to this type of problem using a between-subjects Session × Experimental Condition ANOVA. A main effect of experimental condition emerged, $F(1, 21) = 4.62$, $MSE = 0.197$, $p < .05$. There was no main effect of session ($F < 1$) and no interaction, $F(1, 21) = 1.12$. The main effect of experimental condition reflected fewer extra moves during planning (3.5) than during control sessions (8.2), consistent with our findings for other problems.

Reasons for the higher solution rate among control participants during Session 2 (who had planned in Session 1) were explored by analyzing the concurrent reports of 7 participants who solved the problem and 3 who did not solve it. Participants were selected for analysis on the basis of their accuracy at solving earlier problems: As with our other analyses, at least three other problems had to be correctly solved for inclusion. Including more participants increased the magnitude of observed differences, but we were concerned that differences in solution accuracy on the final problem could mainly reflect ability differences rather than strategic differences.

Verbal reports were coded for evidence of complex planning strategies, but we found only one case where a participant used sophisticated analysis (in a solver). However, we found five separate cases of restarting the problem in the successful participants and two cases in the unsuccessful ones, a pattern possibly indicating continued depth-first search. Five participants who were successful used, on several occasions each, some form of backtracking involving looking ahead one or more steps before moving and then mentally undoing the moves to try something else; none of the unsuccessful participants did so. We calculated the number of instances of looking ahead one or more steps for each participant. A two-sample t test adjusted for unequal variances was not significant but approached significance, $t(6) = 2.24$, $p = .07$, with a trend for solvers to look ahead more frequently (1.86 times/problem) than nonsolvers (0 times). With such a small sample, conclusions must be qualified, but it appears that some of the participants may have continued to use limited look-ahead planning. However, they interspersed planning and action rather than planning the entire sequence in advance. This may explain why their solution rate improved, but nonetheless they produced many more moves than planners.

Summary of Main Results

Planning produced qualitatively different solution processes than allowing participants to solve problems without any special instructions. Rather than making many looping moves, planners relied on depth-first search of the problem space, which minimized working-memory demands by frequently returning to the beginning of the problem. In contrast, control participants generally behaved much like Atwood and Polson's (1976) participants, making many looping moves and consequently very long solution paths. Therefore, we replicated the main results of earlier studies in the Session 1 control group, but planning participants consistently created more efficient solutions than control participants did. Planners were more likely to discover complex planning strategies and there was some weak evidence that, when no longer required to plan, participants may have continued to intersperse planning and action.

Experiment 2

In Experiment 2 we explored whether exposure to planning instructions influences solution even after planning requirements are lifted. Participants solved water jugs problems for two sessions, as in Experiment 1. However, in Experiment 1, participants always received one planning session and one control session. For Experiment 2, to separate the effects of practice and planning, some participants received two sessions of planning and others received two sessions of control instructions in addition to the conditions from Experiment 1.
We also developed a mouse-driven computer interface, which made it easier and faster to make moves, and attempted to increase control over the retrieval of planned solutions. After planning participants generated their solutions, they completed a 90-s filler task designed to prevent rehearsal in STM.

Method

Participants

Forty-five Florida State University undergraduates participated for course credit. Seventeen were excluded on the basis of the arithmetic test used in Experiment 1, and 4 never completed Session 2. Data from the remaining 24 participants were analyzed.

Materials

Two problem sets (Set C and Set D, see Table 1) were generated that were solved 80% of the time or more in pilot studies. Each set consisted of a practice problem and four water-jugs problems. Participants were randomly assigned to solve set C in the first session and set D in the second session, or vice versa. The cover story can be found in the Appendix.

Procedure

Testing occurred in two 1-hr sessions approximately a week apart. As in Experiment 1, participants learned to think aloud, solved the screening problems, and read the cover story. They were led through the practice problem to ensure that they understood the rules, and then they solved the problems following the same procedures outlined in Experiment 1. The only other changes involved attempts to minimize errors introduced by the user interface and calculation errors. A mouse-driven interface was used instead of a keyboard interface. Experimenters monitored participants' verbal reports and warned them if their verbalized move and the move they entered using the interface mismatched. Participants were warned whenever their verbalized jug contents did not sum to the total amount of water available in the problem.

In each of the two sessions, participants were either told to plan their solutions as in Experiment 1 (planning condition) or to make moves whenever they wanted to (control condition). Participants could be in the same condition across both sessions or be in different conditions in each session, resulting in four groups: control/control, control/plan, plan/control, and plan/plan. The filler task in the planning condition involved reading random letters aloud from the screen at a rate of 2/s for a period of 90 s prior to entering the solutions.

Results and Discussion

In all subsequent analyses, Problem Sets C and D had approximately equal difficulty: We found no effects of problem set (C vs. D) and no significant interactions with problem set in each case. In subsequent analyses we used mixed-factorial ANOVAs. The within-subject factors were problem and session (Session 1 vs. Session 2). The between-subjects factors were condition/S1 (planning vs. control during Session 1) and condition/S2 (planning vs. control during Session 2). Post hoc comparisons used Bonferroni corrected two-tailed t tests that compared all pairs of means.

Solution Accuracy

A problem was counted as correctly solved if the participant entered a set of moves that resulted in the goal state before the time limit elapsed. The proportion of correct solutions across all problems was 80%. As in Experiment 1, our primary concern with respect to solution accuracy was whether problem type would interact with experimental condition variables, which would complicate interpretation of other results. Fortunately, problem type was not involved in any higher order interactions. There was only a main effect of problem type, $F(3, 60) = 4.33$, $MSE = 0.130$, $p < .01$, indicating that some problems were solved more frequently than others (92%, 67%, 77%, and 85% for Problems 1 to 4, respectively).

No between-subjects effects of experimental condition (planning vs. control) emerged for either session, and there was no Condition/S1 × Condition/S2 interaction (all Fs < 1). Experimental condition did interact with session: Session × Condition/S1, $F(1, 20) = 4.39$, $MSE = 0.119$, $p < .05$, and Session × Condition/S2, $F(1, 20) = 4.39$, $MSE = 0.119$, $p < .05$, both of which qualified a main effect of session, $F(1, 20) = 6.32$, $MSE = 0.119$, $p < .05$. Post hoc tests revealed that those participants who planned during Session 1 were more accurate during Session 2 (92%) than they were during Session 1 (69%)—in other words, they improved substantially—whereas those who were in the control group during Session 1 had similar solution rates across sessions (79% and 81%). Similarly, those participants who received control instructions during Session 2 were reliably more accurate during Session 2 (94%) than they had been during Session 1 (71%). Those who received planning instructions during Session 2 had similar solution rates across sessions (77% and 79%). There were no other reliable effects.

Number of Moves and Plan Execution Time

As in Experiment 1, we next analyzed the number of moves beyond the minimum required to solve each problem. Values were $\log_{10}(m + 1)$ transformed and missing values were replaced with the grand mean of the transformed values. The Problem × Session × Condition/S1 × Condition/S2 mixed ANOVA uncovered no three-way or four-way interactions, so we reported the between-subjects analyses first, followed by the within-subject analyses.

Our hypothesis was that planning in either session ought to benefit performance relative to control instructions. Consistent with our predictions, significant main effects of condition/S1, $F(1, 20) = 55.47$, $MSE = 0.147$, $p < .001$, and condition/S2, $F(1, 20) = 7.71$, $MSE = 0.147$, $p < .05$, were observed, with the planning condition resulting in shorter solutions than the control condition. There was also a significant interaction between condition/S1 and condition/S2, $F(1, 20) = 7.71$, $MSE = 0.147$, $p < .05$. Post hoc comparisons revealed that control/control participants made more moves beyond the minimum than any of the other three groups, which did not differ from each other.

For the within-subject effects, a significant interaction between session and condition/S2 was found, $F(1,60) =$

5.46, MSE = 0.132, p <.05. Figure 3 shows the means for the planning and control conditions as a function of session. Post hoc comparisons showed that when participants planned during Session 2, they made fewer extra moves (1.43) than they had made in Session 1 (3.45), reflecting shorter overall solutions when people planned. As expected, the planning condition in Session 2 made fewer moves (1.43) than the control condition did in Session 2 (3.74).

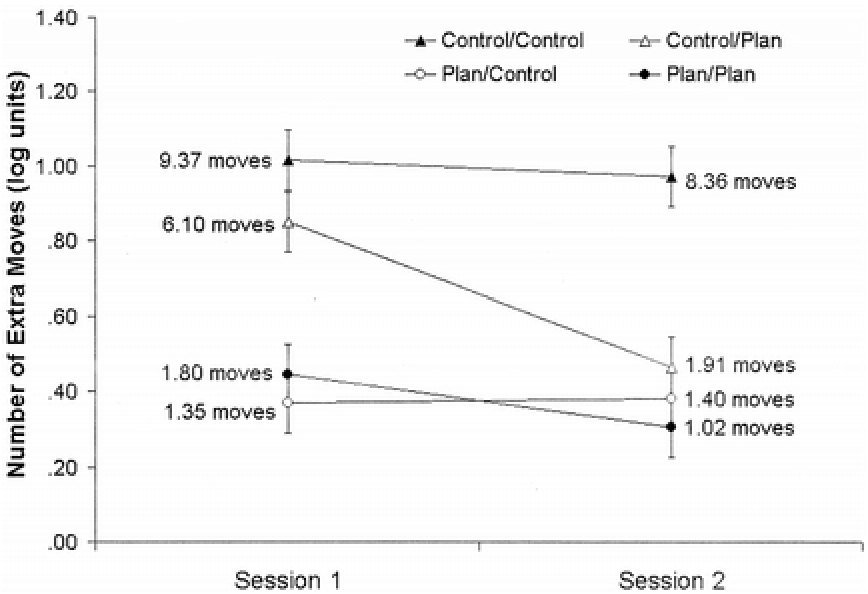

Figure 3. Number of moves committed beyond the minimum required to solve a problem as a function of experimental condition and session in Experiment 2. Axis values were log(x + 1) transformed. Error bars represent plus or minus standard errors

Finally, there was a main effect for problem type, F(3, 60) = 6.51, MSE = 0.183, p <.001. Post hoc tests revealed that Problem Type 2 required more extra moves to solve (5.32) than either Problem Type 1 (1.72) or Problem Type 4 (2.39) but did not differ from Problem Type 3 (3.34). No other significant differences emerged between the problems.

Maintaining plans over a 90-s filled retention interval, intended to prevent maintenance in the articulatory loop, did not prevent effective planning. To verify that participants in the planning condition recalled plans from memory rather than solving the problems again, the time to enter solutions was examined. Because the time to make the first move represents time spent planning and reading-understanding the problem, we compared the solution times without the first move. If planners recalled the plans from memory, then the time to enter the solutions should be much shorter in the planning condition than in the control condition (cf. Phillips, Wynn, McPherson, & Gilhooly, 2001). We conducted an Experimental Condition × Problem ANOVA on the log-transformed solution times, revealing a significant main effect of experimental condition, F(1, 22) = 71.95, MSE = 0.135, p <.001. Session 1 solution times were much shorter for the planning group (32 s) than the control group (143 s). Because entering moves using the interface took about 3 s, planners required only about 3–4 s to retrieve each step from memory.

Because we were interested in practice effects across sessions, separate Session × Problem Type ANOVAs were conducted for each of the four combinations of planning and control instructions (plan/plan, plan/control, control/plan, and control/control) to find out whether performance changed from Session 1 to Session 2. Only the control/plan group showed a significant change, F(1, 5) = 8.37, MSE = 0.215, p <.05, reflecting better performance when participants switched from control instructions in Session 1 to planning instructions in Session 2 (replicating findings from Experiment 1). Interestingly, however, performance in the plan/control group, who switched from planning instructions to control instructions, did not change (F < 1). Thus, participants apparently continued to benefit in Session 2 from a planning instruction delivered during Session 1, whereas participants in the control/control group seemed to learn little, which is consistent with findings showing that problem solving alone often produces little transfer (e.g., Sweller, 1988; Sweller & Levine, 1982).

One plausible explanation for the sustained benefit of planning is that participants continued to plan after finding that it was an effective way to solve problems. Earlier we showed that the time to enter the first move was much longer in the planning condition than in the control condition, suggesting that participants in the planning condition spent the bulk of their time planning rather than entering moves. If participants continued to plan in the control session, we would expect to see a similar pattern. For all correctly solved problems, we therefore calculated the proportion of total solution time spent on the first move for the plan/control group ($M = .70$, $SD = .25$) and the control/control group ($M = .13$, $SD = .08$). The plan/control group was faster than the control/control group, $t(43) = 8.85$, $p < .001$, suggesting that participants in the control group continued to prepare solutions in advance if they had been exposed to planning instructions earlier.

Experiment 3

Experiment 2 examined the benefits of planning during the first session and how exposure to planning instructions improves solutions even after that instruction was removed. Participants in the planning conditions also generated plans for their current problem and maintained them in memory without relying on rehearsal. In Experiment 3 we examined the durability of these planned solutions in memory. It is sometimes useful to learn solutions to particular problems, in case they might be encountered again. Without planning instructions, memory for past solutions on tasks like the Tower of Hanoi (Karat, 1982) and Missionaries and Cannibals (Reed et al., 1974, Experiment 2) is generally poor. That planned solutions survived without rehearsal suggests that planned solutions might be longer lasting. However, remember that planners in Experiment 1 used depth-first search and frequently restarted problems, suggesting that even for them memory may have been a limiting factor. Therefore, in Experiment 3 we assessed memory for previously produced plans by asking participants to recall information about their final two solutions. We also tried to assess whether making problems more distinctive might facilitate memory for previous solutions by giving participants either the same cover story on all problems or introducing a different cover story for each (which was then used at test to cue memory for the problem).

Method

Participants

Fifty-one Florida State University undergraduates participated for course credit, of which 32 met all requirements for analysis. Thirteen failed to pass the arithmetic screening pretest (as in Experiments 1 and 2). Six participants were excluded because they failed to solve one of the two target problems (and hence could not be meaningfully tested on their memory for those solutions).

Design and Materials

The design was 2 (instruction) × 2 (story) × 2 (problem) × 2 (test order). All variables were between subjects except for problem, which was manipulated within subject.

Instruction variable

For the problem-solving portion of the experiment, participants were randomly assigned either to the planning condition, in which they received an instruction to plan similar to that in earlier experiments (the planning condition), or the control condition in which they were simply told to solve the problems and make moves at any time.

Story variable

Participants were also randomly assigned to either a condition where they received potential extra retrieval cues in the form of a new cover story for each problem they solved (the “stories” condition), or a condition in which they saw the same cover story for all problems (the “single-story” condition). The full cover stories are listed in the Appendix.

Problem variable

Participants solved three problems and received a final recall test on the last two, which we called the middle problem and the final problem. The within-subject contrast between the middle problem and the final problem was encoded as the problem variable. All three experimental problems and two practice problems are displayed in Table 1 (Set E). The order of the two target problems, Problems E-2 and E-3, was counterbalanced such that each problem served equally often as the middle or the final problem.

Test-order variable

During the recall of the last two problems, half of the participants recalled the most recently solved problem first followed by the middle problem (last-first condition), whereas others were tested in the opposite order (middle-first condition).

Procedure

The experiment was conducted in a single 90-min session. Prescreening followed the procedure of Experiment 2, but participants solved two practice problems instead of one. The experimental procedures followed those of the first session of Experiment 2, except that the maximum time to solve problems was increased to 15 min, and in the planning condition, participants executed their plan immediately when ready instead of following a filled retention interval. In the stories condition, the new cover stories appeared between problems and provided new names for task elements (e.g., grog barrels became lemonade pitchers). During problem solving, the screen did not reflect these changes and was the same for every problem.

After completing all three problems, participants received a surprise recall test on the last two problems (the target problems). For each problem, participants were asked to recall the sizes of the containers and the goal state for each container. If they could not remember, they were required to give their best guess. They were then asked to recall the solution and required to give at least the first move if they reported they were unable to recall the whole solution. They were instructed to avoid guessing and not to try to solve the problems again; we were able to monitor this by having them continue to think aloud during the testing phase. Whenever participants began doing calculations aloud, we asked if they were solving the problem again and if so they were stopped and no further recall was counted.

Scoring of Postsession Recall

The first of the two dependent measures for the surprise recall test assessed recall for problem elements and equaled the sum of the number of correct goal elements recalled (0–3) and the number of correct container size elements recalled (0–3). Goal elements and container sizes had to be in the correct location to be scored as correct. For example, if the goal was 6/0/3, the jug sizes were 9/5/4, and the goal was recalled in scrambled order as 6/3/0 but the jug sizes were correct (9/5/4), the recall measure would be 4. The score was converted to a proportion.

The second dependent measure assessed recall for correct solution steps starting from the first recalled move. The structure of the problem space affords several possible correct solutions. The three solution types were: (a) Type M solutions: the minimum path solution; (b) Type 2 + M solutions: two moves followed by the minimum path; or (c) Type L solutions: an alternate, much longer, but nonetheless correct and nonlooping solution path. Depending on which type of solution a participant found, we counted the number of correct steps recalled from their solution path and divided by the number of steps along the relevant solution path to yield a proportion recall. This was based on the minimum path to the solution not the total number of moves they committed. Thus, a participant who made 35 moves could still have recalled only the portion of their solution that corresponded to their minimum path and received full credit. Participants received no credit for recalling steps that led to looping moves.

Results

Analysis of Initial Solution Characteristics

Only data from participants who solved both target problems successfully were analyzed (16% of participants failed to solve Problem 2 or Problem 3). Problem 1 was solved correctly by over 90% of participants in all conditions.

Our first analysis checked whether we replicated the benefits of planning on number of moves seen in our previous experiments. For the analyses of number of moves beyond the minimum, we used a mixed-design ANOVA with one repeated factor, problem, and two between-subjects factors, instruction (plan or control) and story (same story throughout the experiment or different story for each problem). The dependent measure was log10(m + 1) transformed. The critical result was that planning participants made fewer moves beyond the minimum (1.87) than control participants (15.4), $F(1, 28) = 104.15$, $MSE = 0.025$, $p < .001$. No main effects of story ($F < 1$), problem, $F(2, 56) = 1.25$, $MSE = 0.025$, or interactions emerged.

Postsession Free Recall of Problem Elements and Solution Steps

Separate analyses were conducted on the proportion of solution steps recalled and the proportion of problem elements recalled. For each dependent measure, we used an Instruction Condition × Story × Problem × Test Order mixed ANOVA (details on independent variables are in the design section).

Memory for solution steps

Solutions were classified into the three types described in the method section (L, M, and 2 + M). Only 2 participants in the planning group discovered the Type L solution and none in the control group discovered it. The remaining solutions were evenly split between M and 2 + M solutions.

Table 3 shows proportion of solution steps recalled. A main effect of instruction emerged, showing better overall recall in the planning group than in the control group, $F(1, 24) = 34.35$, $MSE = 0.047$, $p < .001$. There was also a main effect of problem, $F(1, 24) = 7.52$, $MSE = 0.071$, $p < .01$, reflecting a memory advantage for the final problem. However, the results were qualified by a significant Problem × Instruction interaction, $F(1, 24) = 5.90$, $MSE = 0.071$, $p < .01$. We followed up the interaction using Bonferroni-corrected t tests. Our main interest was whether planners would recall more solution steps for each of the two target problems than control participants. The planning group recalled more steps on the final problem than the control group (78% vs. 20%), but they did not differ on the middle problem.

Table 3
Proportion of Solution Steps Recalled in Experiment 3

Problem/test order	Planning				Control			
	1 story		3 story		1 story		3 story	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2/middle first	.30	.26	.25	.38	.20	.00	.05	.10
2/final first	.50	.42	.40	.43	.00	.00	.45	.30
3/middle first	.65	.34	.65	.47	.30	.26	.10	.20
3/final first	1.00	.00	.80	.23	.00	.00	.40	.43

Note. Per-cell $n = 4$.

Proportion of Solution Steps Recalled in Experiment 3

In addition, a triple interaction emerged among story, instruction, and test order, $F(1, 24) = 7.76$, $MSE = 0.047$, $p < .01$. To interpret this interaction within the framework of our interest in planning, we conducted separate Instruction × Story ANOVAs at each level of test order. When the last problem was tested before the middle problem, there was only a main effect of instruction, reflecting significantly better recall of solutions in the planning condition than the control condition, $F(1, 28) = 7.78$, $MSE = 0.093$, $p < .01$. When the middle problem was tested before the last problem, there was a main effect of instruction, $F(1, 28) = 16.55$, $MSE = 0.103$, $p < .001$, and an interaction between the story and instruction, $F(1, 28) = 6.40$, $MSE = 0.103$, $p < .05$. For the single-story condition, the planning group showed better recall than the control group, $t(14) = 5.56$, $p < .001$; however, in the multiple-story condition, there was no difference between the planning and the control groups ($t < 1$). In summary, we found consistently superior memory for planned solutions, except in the condition where participants received a new story for each problem and were tested on the middle problem before the last problem.

Memory for problem elements

The same Instruction × Story × Problem × Test Order mixed ANOVA was used to analyze the proportion of problem elements recalled. The planning group recalled a larger proportion of problem elements than the control group, as reflected in the main effect of instruction, $F(1, 24) = 5.24$, $MSE = 0.101$, $p < .05$. There was also a main effect of test order, $F(1, 24) = 20.49$, $MSE = 0.084$, $p < .001$. However, there was a three-way interaction among problem, test order, and instruction, $F(1, 24) = 4.34$, $MSE = 0.084$, $p < .05$ (see Table 4). There were no other significant main effects or interactions.

Table 4
Proportion of Problem Elements Recalled in Experiment 3

Problem/test order	Planning				Control			
	1 story		3 story		1 story		3 story	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2/middle first	.63	.25	.50	.41	.54	.34	.21	.08
2/final first	.83	.33	.50	.58	.42	.17	.13	.16
3/middle first	.80	.17	.75	.29	.58	.48	.50	.36
3/final first	1.00	.00	.83	.34	1.00	.00	.96	.08

Note. Per-cell $n = 4$.

Proportion of Problem Elements Recalled in Experiment 3

We followed up the three-way interaction with separate Instruction × Test Order ANOVAs for each problem. For the middle problem, there was only a main effect of instruction, with better memory for problem elements in the planning condition, $F(1, 28) = 4.55$, $MSE = 0.110$, $p < .01$. The main effect of test order and the Instruction × Test Order interaction were not significant (both $Fs < 1$). For the final problem, however, there was only a main effect of test order, such that participants recalled more when the problem was tested first than when it was tested second, $F(1, 28) = 4.74$, $MSE = 0.081$, $p < .05$. Neither the main effect of instruction nor the interaction was significant. In summary, planners had better memory for the elements of the middle problem but not for the most recent problem, perhaps because of the overall high recall rates for the final problem.
Last, we expected that multiple stories might improve memory by adding retrieval cues. There was a main effect of story, but contrary to our prediction, participants who received a new story for each problem recalled a smaller proportion of problem elements (.53) than participants presented with the same cover story for all problems (.73), $F(1, 24) = 6.51$, $MSE = 0.101$, $p < .05$.

Discussion

As in Experiments 1 and 2, participants in the planning condition generated solutions with fewer moves than control participants did. Memory for problem elements was good immediately after solving a problem whether participants planned or not, but apparently planners encoded the problem elements in a way that protected them from interference, whereas control participants rapidly forgot them.

The results differed for solution steps. Planners showed almost perfect recall for the most recently solved problem when it was tested first. However, when the most recently solved problem was tested second, recall was lower. Memory for earlier solutions was also rather low. In contrast, the control group showed uniformly poor memory for their solutions regardless of test order. In summary, participants in the planning group, but not the control group, could reproduce some of their earlier solutions.

These results suggest that memory for solutions was reduced by RI generated during later recall of solutions or later problem solving. Although in the absence of interference the plans remained active for a long period of time, consistent with storage in LTM, the impact of RI on memory for plans emulates some characteristics of classic STM. Specifically, plans cannot be fully retrieved after solving more problems or after recalling other solutions. The presence of severe RI might explain why planners in Experiment 1 frequently restarted the problem and used depth-first search: Near-perfect recall was available only for the most recently generated solution.

Presenting new cover stories for each problem and using the stories as retrieval cues did not improve recall of solution plans and, in fact, resulted in poorer recall of the initial and goal states. Perhaps the information in each cover story increased the total amount of information to be recalled, thereby reducing access to information about the problem. Most participants may have ignored the stories and simply referred to the containers by letter or size rather than as, for example, pitchers or test tubes, given the fact that the cover story did not alter the display. Story names may not have served as an effective retrieval cue for the solution because participants may have encoded the problems on the basis of order and not the story.

In summary, participants in both conditions showed substantial long-lasting memory for recently encountered problems but relatively poor memory for solutions, with the primary exception being the planners' accurate immediate memory for the most recently generated solution.

Experiment 4

Our first three experiments showed that only planners maintain virtually perfect memory for their most recent solution plan. However, Experiment 3 revealed that planners' memory for solutions declined dramatically after solving additional problems (or even recalling a different solution). Furthermore, we suggested that it was RI rather than storage in an STM or working memory that led to the rapid forgetting, which is plausible because the tasks that caused forgetting all involved processing the same type of stimuli, namely, solutions to water jugs problems. In Experiment 4, we unexpectedly tested participants' memory for planned solutions to water jugs problems after a filled retention interval. The retention interval was filled with tasks that varied in their similarity to the original task. If processing of similar material were the critical factor, one would predict that interpolated activities unrelated to water jugs solutions would lead to less interference for memory for solution plans. In contrast, if the plans were never stored in LTM, they would become inaccessible regardless of the filler tasks' content.

Method

Participants

Seventy-one University of Florida and Florida State University undergraduates participated for course credit. Nine participants were replaced because of the prescreening for arithmetic skill (as in Experiments 1–3). Twelve participants failed to solve one of the two required water jugs problems (Problem 2 or Problem 3). The final sample consisted of 50 participants whose data were submitted for analyses.

Problems

The water jugs problems were identical to those used in Experiment 3. The Tower of Stockholm problem is an isomorph of the Tower of Hanoi (e.g., Karat, 1982; Kotovsky, Hayes, & Simon, 1985) consisting of three pegs and three discs. The discs were of the same size but were labeled with the numbers 8, 5, and 3 and stacked on the leftmost peg in order from the largest to the smallest. The object of the task was to get all three discs on the rightmost peg, following the rule that a disc cannot be placed on a disc with a smaller number. The participant could only move one disc at a time and could only move the top disc on a peg.

Procedure

The procedure followed that of Experiment 3. As before, participants solved Problem E-1 from Table 1 and then a target problem (either Problem E-2 or E-3 from Table 1). The only thing that varied across the three experimental groups and one reference group was the type of filler task that followed the target problem. The reference group was the two-jug condition ($n = 12$) in which the surprise recall test was administered immediately. We included this group to show that initial recall levels were at ceiling before RI was induced through filled retention intervals. In the three-jug condition ($n = 14$), the filler task was another water jugs problem solved in the same fashion as the previous two and was thus expected to be maximally similar. This condition corresponded to the planning group from Experiment 3. In the tower condition (n = 12), the filler task involved reading the instructions for the Tower of Stockholm task and producing and executing a planned solution to that problem. Participants were told that the problem could be solved in exactly seven moves. This task was intended to be of intermediate similarity to the target water jugs problem. Finally, in the reading condition (n = 12), the participant read aloud part of a biography of Martin Luther King, expecting a later comprehension test. Each participant's reading time was matched to a solution times for a three-jug condition participant. To ensure that participants comprehended the text, they summarized it aloud at the end of the experiment. The surprise recall test was the same as in Experiment 3, except that only the solutions and not the problem elements were tested.

Results and Discussion

Participants who correctly solved Problem 2 (and, in the three-jug and tower conditions, Problem 3) solved Problem 1 96% of the time, with a mean number of moves beyond the minimum of 0.31 (SD = 1.01). The mean number of moves beyond the minimum for the target problem was 0.72 (SD = 2.57). All participants in the reading condition were able to produce a summary of the text. The mean total time spent on the third jugs problem and the reading was 7 min, 1 s (these conditions were matched for time). For the third jugs problem, the mean number of moves beyond the minimum was 0.71 (SD = 1.68). For the Tower of Stockholm task, it took an average of 1 min, 53 s to solve the problem plus 2 min, 30 s to learn the rules. All participants solved the tower task in the minimum number of moves.

Our central question was how memory for the planned solution to the target problem would vary as a function of the filler task. The dependent measure was the proportion of correct solution steps recalled, which was calculated in the same fashion as in Experiment 3. A one-way ANOVA on the proportion of solution steps recalled by experimental condition (excluding the two-jugs group) showed that there was a main effect of condition, $F(2, 35) = 8.43$, $MSE = 0.130$, $p < .001$. Post hoc comparisons using studentized Newman-Keuls revealed that the three-jugs group recalled fewer steps than the tower group, which in turn recalled fewer steps than the reading group (see Figure 4). All participants in the two-jug group recalled the solution to the target problem without error.

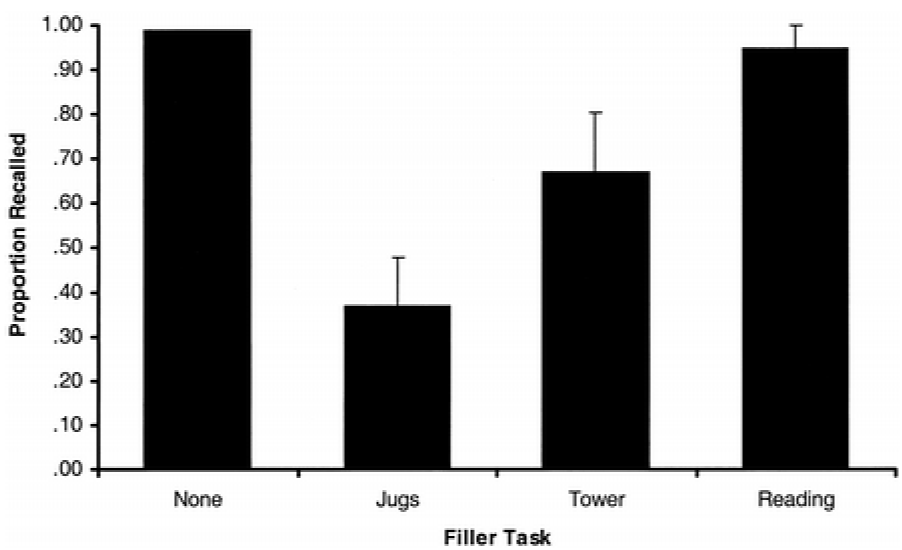

Figure 4. Proportion of solution steps from the target problem recalled correctly as a function of the filler task in Experiment 4. Error bars represent plus or minus standard errors

Memory for planned solutions depended on the type of interpolated task. As might be expected from RI, memory for the solution of the target problem was reduced both by solving a nearly identical problem (the three-jug condition) and by solving a moderately similar planning problem that shared common elements (the tower condition). Admittedly, the locus of interference could be either from solving similar problems or from retrieving them (or both). We were unable to detect any forgetting caused by a very dissimilar task (reading text), consistent with our previous findings that the most recent plan can be maintained over a long period of time in the absence of similar interfering information.

In summary, executed plans were available for recall following long filled retention intervals (even when tested unexpectedly). However, they were subject to RI from solving other problems. Complete planned solutions, although likely stored in LTM during planning or execution, were forgotten following additional similar problem solving. Thus, planning would be unlikely to aid in learning individual solutions unless later activities cause little RI.

The RI induced by processing solutions to water jug problems provides a mechanism explaining why planning participants use depth-first search and frequently restart problems. Keeping a solution path active requires ensuring that few (if any) alternative solutions are considered after generating an almost-correct path. Otherwise, memory for the solution path is subject to substantial RI from the later-considered paths.

General Discussion

Typical descriptions of novice problem solving focus on how people use local information combined with heuristics such as hill climbing to make choices during problem solving (e.g., Altmann & Trafton, 2002; Atwood & Polson, 1976; Newell & Simon, 1972; Simon, 1975). People also use means-ends analysis and subgoaling to break problems down into subparts (e.g., Anderson & Douglas, 2001; Newell & Simon, 1972) and their history of success to guide future choices, combining it with the other aforementioned strategies (e.g., Lovett & Anderson, 1996; Lovett & Schunn, 1999; Schunn, Lovett, & Reder, 2001). Other types of planning strategies have received relatively little attention in the study of novice problem solving. Perhaps this is because subgoaling, hill climbing, and history of success with particular operators are sufficient to explain the behavior of the majority of participants in tasks such as river-crossing problems (Jeffries et al., 1977), water jugs (Atwood et al., 1980; Atwood & Polson, 1976), and isomorphs of water jugs (Lovett & Anderson, 1996). We first discuss the kinds of changes to problem solving that occur in response to planning instructions, and then we discuss why people usually do not plan in unfamiliar tasks.

Effects of Planning Instructions

Our research identified ways that planning instructions influenced memory during problem solving (and after). Whereas memory for the descriptions of the presented problems was good for both planning and control participants' most recent problem, memory for earlier problem descriptions was reduced less for the planning participants than for control participants. Planners' immediate memory showed a strong recency effect, and their memory for their most recent solution was almost perfect. Control participants had poor memory even for their most recent solution. Furthermore, planners relied on recency in generating their solutions by using depth-first search and restarting the problem and they were able to avoid the usual errors made in the control group (Experiment 1). Finally, processing of similar information prior to recall of the solution reduced recall for water jug solutions, consistent with RI (Experiment 4). Thus, a major challenge for planners in an unfamiliar task is likely RI produced from problem solving and consideration of many similar alternative courses of action.
In addition, we found evidence that planning improves the solutions produced through planning (short-term changes) and changes how future problems are approached (long-term changes). The short-term changes may be attributable to an increased use of reflective planning strategies that allowed extensive look-ahead planning (Experiment 1). An additional short-term change was that planning includes a de facto requirement to memorize solutions. Although control participants frequently could not reproduce even their first solution step, planners often had a good representation of their solution after planning and executing it (Experiments 3 and 4).

Instructions to plan changed participants' task representations in ways that maximized efficiency without generally reducing solution accuracy. In many real-world tasks, reducing error rates is more important than total time because the cost of errors is quite high (e.g., defusing bombs). Although simply presenting problems for solution may lead to the development of the kinds of mental representations needed to reduce errors, goal-directed problem solving alone often leads to minimal improvement or transfer (Sweller, 1988), something also found in our Experiment 2 control group.

Although our planning manipulation did not generally improve solution accuracy (with the possible exception of the hardest problem in Experiment 1), our results nonetheless can be related to the beginning of the development of expertise in other problem-solving activities with similar task properties (e.g., chess). Experts whose superior planning and improved solution accuracy develop after extended practice show remarkably good memory for specific states, such as chess positions encountered in a move selection task, demonstrating that they find ways ultimately to cope with the severe interference effects observed in our study (Ericsson & Kintsch, 1995). We believe that when individuals are motivated to plan, they are also motivated to find more distinctive state representations to reduce interference. This is why when experts plan and decide, they rely on depth-first search strategies (e.g., Charness, 1981) and higher level representations (Saariluoma, 1995), just as our participants did. In our short study, only a few participants developed such higher level representations, and their impact on accuracy emerged only on the hardest problems. Our results nonetheless suggest that inducing planning might ultimately lead to task representations that support superior memory for states and thus better reasoning, as they do in chess.

It is worth trying to relate known methods for improving solution efficiency to planning in the water jugs, both to determine which mechanisms might be involved and to guide the development of effective training methods. First, teaching solutions to analogous problems to produce explicit transfer of recalled solutions can sometimes reduce errors (e.g., Bassok, 1990; Ross, 1987, 1989; VanLehn, 1988). Unfortunately, in our water jugs problems, participants had relatively poor memory of earlier solutions because of extreme RI between successive solutions (Experiments 3 and 4). Second, requiring participants to explain the reasons for their move selection is also associated with greater transfer of problem solving (Berardi-Coletta, Buyer, Dominowski, & Rellinger, 1995; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Davies, 2000; Gagne & Smith, 1962). The benefits of planning could be related to explanation giving, because planners used abstract reasoning strategies more frequently than control participants. Third, enhanced transfer and learning are obtainable when exploration of the problem space rather than achieving a particular goal is encouraged (e.g., Sweller, 1983). Perhaps planning helped participants learn move patterns, resulting in more efficient solutions, as demonstrated in Experiment 2, where participants made better move choices in a control condition after having been required to plan during earlier sessions.

Perhaps the simplest possibility is that participants transferred the strategy of searching ahead a few moves. Consistent with continued planning in some participants, in Session 2 of Experiment 2, control participants who had been exposed to planning instructions earlier deliberated longer before making their first move than participants who had never been exposed to planning instructions, probably because some participants continued to plan even after planning was no longer required.

In summary, planning instructions changed problem solving in a number of ways. Although RI seemed to make it difficult to retain planned solutions to several similar problems, planning instructions may lead to lasting changes in mental representations of the task, especially when a problem is difficult to solve using the typical heuristic approaches to problem solving.

Planning in Unfamiliar Tasks: Why or Why Not?

Although our participants readily transferred the benefits of planning after exposure, people typically do not plan in unfamiliar tasks. One might reasonably wonder why not. Previous accounts focused on memory capacity limits (Atwood et al., 1980; Atwood & Polson, 1976; Jeffries et al., 1977; Simon & Reed, 1976). Memory capacity-based accounts have also been proposed for other problem-solving effects, including why some problems are more difficult to solve than others (e.g., Kotovsky et al., 1985). Our findings seem difficult to explain through limited memory capacity alone. Participants in both the planning and the control conditions recalled starting and goal states from previously solved problems well, especially for the most recently solved problem. Hence, participants in the control condition only showed poor memory for solutions. Second, participants generated and recalled plans mentally despite extensive working-memory demands incurred by our restrictive planning procedure. During the plan-generation phase, planners generated all intermediate states of their solutions without external memory support. Planning participants executed a correct solution plan to most problems, implying that they mentally constructed (and kept in working memory) information about the current state that would have been available on-screen in the control condition. Taken together, our findings suggest that a complete explanation of participants' reasons for choosing not to plan should consider not just memory load but the representations participants use to cope with interference and history of success with particular problem-solving strategies. Other recent work has begun to suggest representation-driven alternatives to memory capacity explanations in other problem-solving domains, and we view our work as concordant with those accounts (Clement & Richard, 1997).

Our belief is that people may approach novel problems without planning because complete look-ahead planning is often unnecessary to solve everyday problems. In everyday life, hill climbing and means-ends analysis are useful tools for solving our problems. However, puzzles often are difficult precisely because our usual heuristics are inapplicable or even misleading (Payne, 2001; Reitman, 1973). In water jugs, for example, hill climbing often leads to irreversible looping moves away from the goal (Atwood & Polson, 1976). Similarly, Atwood and Polson (1976) suggested that trying to identify subgoals a priori is impossible in our water jugs task. It makes sense that people would first try what has worked on other problems in the past. What is surprising about our results is that people do not avoid planning once they have tried it, even though they rarely discover it spontaneously. We envision two plausible accounts of this phenomenon that are not mutually exclusive.

The active-procedure bias account suggests that people tend not to seek alternative approaches to a problem in many circumstances. The recent represent, construct, choose, and learn (RCCL) model of problem solving, for example, implies that people will not seek alternative representations of the problem unless the current representation fails to produce obvious progress (Lovett & Schunn, 1999). This viewpoint is consistent with classic results on the mental set using an easier version of the water jugs task that showed that people would not usually seek new strategies if the current approach worked (e.g., Luchins, 1942). Even if a strategy is not very effective, people nonetheless may not pause to evaluate their own performance. They may be reluctant to interrupt an active process to engage in reflection (Ellis & Siegler, 1997; Sahakyan, Delaney, & Kelley, 2004). The perceived cost of continuing with an activity that is not clearly failing is small. Once participants have already discovered the option of planning, the likelihood of transferring that discovery and continuing planning increases because planning is known to be effective.

The rational analysis account suggests that people choose the activity least likely to result in substantial cost, where cost depends on time and effort. With the possible exception of the hardest problems, solution rates did not increase when participants planned. People may be aware that incomplete understanding of the rules or calculation mistakes can produce invalid problem states. Additional planning after an error is wasted effort (Anderson, 1990). It might be rational to avoid extensive planning if the costs of planning errors are estimated to be greater than the cost of errors introduced by not planning. As people gain planning experience, they may become more confident or accurate, leading to spontaneous planning. Alternatively, after the frustrating consequences of looping moves when not planning become apparent, planning may begin to seem the less effortful activity, at least for water jugs problems.

Footnotes

1 We are aware that the problem-type variable is difficult to interpret because order of presentation of problems was fixed and, therefore, confounded with the manipulation of number of moves to solution. Another difficulty is that Set A and Set B problems, though equated for a number of solution characteristics, were not absolutely identical. We have included the problem-type variable in the analyses because we feel that it primarily represents the difficulty of the various problems, and interactions between problem difficulty and other variables may be of interest to some readers.

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. Cognitive Science, 26, 39–83.
Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.
Anderson, J. R., & Douglass, S. D. (2001). Tower of Hanoi: Evidence for the cost of goal retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1331–1346.
Atwood, M. E., Masson, M. E. J., & Polson, P. G. (1980). Further explorations with a process model for water jug problems. Memory and Cognition, 8, 182–192.
Atwood, M. E., & Polson, P. G. (1976). A process model for water jug problems. Cognitive Psychology, 8, 191–216.
Bassok, M. (1990). Transfer of domain-specific problem-solving procedures. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 522–533.
Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., & Rellinger, E. R. (1995). Metacognition and problem solving: A process-oriented approach. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 205–223.
Best, J. (1980). Intended and actual moves on river-crossing problems. Psychological Reports, 47, 31–38.
Charness, N. (1981). Search in chess: Age and skill differences. Journal of Experimental Psychology: Human Perception and Performance, 7, 467–476.
Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145–182.
Clement, E., & Richard, J. (1997). Knowledge of domain effects in problem representation: The case of Tower of Hanoi isomorphs. Thinking and Reasoning, 3, 133–157.
Davies, S. P. (2000). Move evaluation as a predictor and moderator of success in solutions to well-structured problems. Quarterly Journal of Experimental Psychology: Human Experimental Performance, 53A, 1186–1201.
Ellis, S., & Siegler, R. S. (1997). Planning as a strategy choice, or why don't children plan when they should?. In S. L.Friedman & E. K.Scholnick (Eds.), The developmental psychology of planning (pp. 183–208). Mahwah, NJ: Erlbaum.
Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. Psychological Review, 102, 211–245.
Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.
Friedman, S. L., & Scholnick, E. K. (1997). An evolving “blueprint” for planning: Psychological requirements, task characteristics, and social-cultural influences. In S. L.Friedman & E. K.Scholnick (Eds.), The developmental psychology of planning (pp. 3–22). Mahwah, NJ: Erlbaum.
Gagne, R. M., & Smith, E. C. (1962). A study of the effects of verbalization on problem solving. Journal of Experimental Psychology, 63, 12–18.
Jeffries, R., Polson, P. G., Razran, L., & Atwood, M. E. (1977). A process model for missionaries-cannibals and other river-crossing problems. Cognitive Psychology, 9, 412–440.
Karat, J. (1982). A model of problem solving with incomplete constraint knowledge. Cognitive Psychology, 14, 538–559.
Klahr, D., & Robinson, M. (1981). Formal assessment of problem solving and planning processes in children. Cognitive Psychology, 13, 113–148.
Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from tower of Hanoi. Cognitive Psychology, 17, 248–294.
Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. Cognitive Psychology, 31, 168–217.
Lovett, M. C., & Schunn, C. D. (1999). Task representations, strategy variability, and base-rate neglect. Journal of Experimental Psychology: General, 128, 107–130.
Luchins, A. S. (1942). Mechanization in problem solving. Psychological Monographs, 54(Whole No. 248).
Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Payne, S. J. (2001). Psychology of everyday problem solving. In N. J.Smelser & P. B.Baltes (Eds.), International encyclopedia of the social & behavioral sciences (pp. 12130–12133). Amsterdam, The Netherlands: Elsevier.
Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the Tower of London task. Quarterly Journal of Experimental Psychology, 54A, 579–597.

Reed, S. K., Ernst, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. Cognitive Psychology, 6, 436–450.
Reitman, W. (1973). Problem solving, comprehension, and memory. In G. J.Dalenoort (Ed.), Process models for psychology (pp. 50–77). Rotterdam, The Netherlands: Rotterdam University Press.
Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 629–639.
Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier examples. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 456–468.
Saariluoma, P. (1995). Chess players' thinking. London, England: Routledge.
Sahakyan, L., Delaney, P. F., & Kelley, C. M. (2004). Self-evaluation as a moderating factor in strategy change in directed forgetting benefits. Psychonomic Bulletin & Review, 11, 131–136.
Schunn, C. D., Lovett, M. C., & Reder, L. M. (2001). Awareness and working memory in strategy adaptivity. Memory & Cognition, 29, 254–266.
Simon, H. A. (1975). The functional equivalence of problem solving skills. Cognitive Psychology, 7, 268–288.
Simon, H. A., & Reed, S. K. (1976). Modeling strategy shifts in a problem-solving task. Cognitive Psychology, 8, 86–97.
Sweller, J. (1983). Control mechanisms in problem solving. Memory & Cognition, 11, 32–40.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257–285.
Sweller, J., & Levin, M. (1982). Effects of goal specificity on means-ends analysis and learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 463–474.
VanLehn, K. (1988). Analogy events: How examples are used during problem solving. Cognitive Science, 22, 347–388.
Ward, G., & Allport, A. (1997). Planning and problem-solving using the five-disc Tower of London task. Quarterly Journal of Experimental Psychology, 50A, 49–78.

APPENDIX

APPENDIX A: Cover Stories

Somewhere in Ethiopia, a young apprentice weaver named Mary has recently become the student of the famous Mistress Zhu, a talented but tough teacher. Zhu has three clay jugs, each of a different size. Each of the jugs has the number of liters of water it can hold written on the side. Last night, Zhu left the largest of the pots outside and it filled with rainwater. Today, she wants Mary to divide the water up a certain way among the jugs for day's work. The problem is, Mary isn't allowed to just eyeball the right amount; she has to measure it exactly. Since the jugs are just plain brown pots, she can't measure the water except by pouring from one jug to another, stopping when the source jug is empty or the destination jug is full. Since she knows how much water each jug can hold, she can figure out how much water will be in each jug after pouring. For example, if she had a jug that could hold 8 liters and was full, and poured into an empty 5-liter jug, then the 5-liter jug would be full and there would be 3 liters left in the 8-liter jug.

Somewhere in the frozen north, a Viking apprentice weaver named Erik has recently become the student of the famous Master Sven, a talented but tough teacher. Sven has three clay pitchers, each of a different size. The pitchers have the number of gallons of water each can hold written on them, but they aren't marked in any other way.

Last night he left the largest one outside and it filled with water from the snow. Today he wants Erik to divide up the water a certain way among the pitchers for the day's work. The problem is, Erik isn't able to just eyeball the right amounts; he has to measure exactly. Because the pitchers are just plain brown vessels, he can't measure the water except by pouring from one pitcher to another, stopping when the source pitcher is empty or the destination pitcher is full. He knows how much each pitcher can hold, so he can figure out how much water will be in each pitcher after pouring. For example, if he had a full 8-gallon pitcher and poured into an empty 5-gallon pitcher, there would be 3 gallons left in the big pitcher, while the 5-gallon pitcher would be full.

Vikings and Grog: Somewhere in the frozen north, an extremely strong Viking named Erik has recently become the student of the master brewer Sven. Sven has three huge barrels, each of a different size. The barrels have the number of Viking mugs of grog each can hold, but aren't marked in any other way. Last night, he finished brewing the grog in the largest barrel, which is now full. Today, Erik has to divide up the grog in a certain way among the three barrels for the day's sales. The problem is, Erik isn't able to just eyeball the right amounts; he has to measure exactly—or else! Vikings are VERY serious about their grog. There's no way to measure grog except by pouring from one barrel to another and adding or subtracting the grog levels. To make sure that the amounts are exact, he can only pour until he empties the barrel he's pouring from or until he fills the barrel he's pouring into. Because Erik knows how much each barrel can hold, he can figure out how much grog is in each barrel after pouring. For example, if he had a full 8-mug barrel and poured into an empty 5-mug barrel, there would be 3 mugs of grog left in the big barrel, while the 5-mug barrel would be full.

Chemists and Chemicals: Dr. Rashid is a famous chemist who is mixing a dangerous, explosive chemical in her lab. Unfortunately last night Dr. Smith's lab chimpanzee from the biology department got loose and broke all of the test tubes except for three. Dr. Rashid knows how big each tube is, and she needs to get EXACTLY the right amount of chemical in each tube for her experiments—or else *BOOM*. The rules are the same as in the previous problem(s), but this time you are working with test tubes and chemicals.

Children and Lemonade: Three business-minded sisters named Sandra, Luna and Mona have opened three lemonade stands together using their dad's famous recipe. They have three big glass pitchers to hold lemonade and they know many SuperTall Cups of lemonade each pitcher can hold. Though they plan to split the profits equally, they know that the three corners they'll be covering get different amounts of business. They want to make sure they have the right number of SuperTall Cups for each corner. The problem is, the girls have to get EXACTLY the right amount of lemonade for each corner or someone will get a glass with less lemonade in it, which is totally unacceptable to the girls. The rules are the same as in the previous problem(s), but this time you are working with pitchers and lemonade.